Apprentissage de réseaux bayésiens hiérarchiques latents pour les études d'association pangénomiques

Raphaël Mourad* — **Christine Sinoquet** ** — **Philippe Leray** *

* *LINA, UMR CNRS 6241, Ecole Polytechnique de l'Université de Nantes, rue Christian Pauc, BP 50609, 44306 Nantes Cedex 3, France {raphael.mourad,philippe.leray}@univ-nantes.fr*
** *LINA, UMR CNRS 6241, Université de Nantes, 2 rue de la Houssinière, BP 92208, 44322 Nantes Cedex 3, France christine.sinoquet@univ-nantes.fr*

RÉSUMÉ. *Nous décrivons un nouveau modèle graphique probabiliste dédié à la modélisation des dépendances statistiques entre marqueurs génétiques du génome humain. Plus précisément, notre travail repose sur les forêts de modèles hiérarchiques latents. L'objectif est de réduire la dimension des données de façon à réaliser d'ultérieurs tests d'association avec la variable indicatrice sain/malade. A cette fin, un algorithme a été développé pour l'apprentissage conjoint de la structure de la forêt et de ses paramètres. Une première implémentation a montré que l'algorithme supporte le passage à l'échelle:* 10^5 *variables pour* 2000 *individus.*

ABSTRACT. *We propose a new hierarchical latent class model devoted to represent statistical dependencies between genetic markers, in the human genome. Our proposal relies on a forest of hierarchical latent class models. The motivation is the reduction of dimension of the data to be further submitted to statistical association tests with respect to diseased/non diseased status. An algorithm, CFHLC, has been designed to tackle the learning of both forest structure and probability distributions. A first implementation has been shown to be tractable on benchmarks describing* 10^5 *variables for* 2000 *individuals.*

MOTS-CLÉS : *Réseaux bayésiens, modèles hiérarchiques latents, réduction de la dimension de données, modélisation des dépendances entre marqueurs génétiques*

KEYWORDS: *Bayesian networks, hierarchical latent class model, data dimensionality reduction, genetic marker dependency modelling*

1. Introduction

Les études d'association génétique sont développées dans le but d'identifier les gènes mises en cause dans les maladies génétiques complexes. La chute des coûts de génotypage permet désormais la génération de centaines de milliers de marqueurs génétiques, plus particulièrement de SNP (Single Nucleotide Polymorphism), couvrant entièrement le génome. Appliquées à l'analyse de populations de patients atteints et de patients sains, les nouvelles technologies de génotypage permettent d'identifier des combinaisons de marqueurs fortement dépendants avec la variable indicatrice sain/malade. Dans ce contexte, les études d'association pangénomiques (*i.e.* couvrant le génome) ont émergé.

L'analyse de ces données génétiques de grande dimension est complexe. La simple recherche d'association entre chaque marqueur et la variable indicatrice sain/malade se révèle difficile car nécessitant des centaines de milliers de tests d'association. Plus inquiétant, les combinaisons de marqueurs génétiques et certains facteurs environnementaux jouent un rôle dans l'apparition de la maladie. De ce fait, un fort taux de faux positifs ainsi qu'une diminution sensible de la puissance statistique, sans parler de la diffulté à traiter les données, représentent de lourds handicaps à surmonter.

Afin de réduire la dimension des données, une idée prometteuse consiste à exploiter l'existence de dépendances statistiques entre SNP, appelées déséquilibre de liaison (LD). Dans le génome humain, le LD est structuré en blocs haplotypiques : des régions de fortes dépendances entre marqueurs sont séparées par de courtes régions présentant de faibles dépendances entre marqueurs. Partant de ce constat, différentes approches ont été proposées telles que les méthodes fondées sur l'inférence d'haplotypes (*i.e.* inférer des données sous-jacentes aux données génotypiques) (Schaid, 2004), sur les blocs haplotypiques (*i.e.* partitionnement des variables haplotypiques) (Pattaro *et al.*, 2008) et sur la sélection de tags SNP (*i.e.* SNP capturant un maximum d'information) (Stram, 2004). Malheureusement, ces méthodes ne prennent pas en compte toutes les dépendances entre marqueurs génétiques, notamment les dépendances entre les blocs haplotypiques.

Les modèles graphiques probabilistes offrent un cadre d'étude adapté pour l'analyse fine des dépendances entre les marqueurs génétiques. Dans cette perspective, plusieurs modèles ont été conçus : principalement, des champs de Markov (Verzilli *et al.*, 2006) et des réseaux bayésiens, avec l'utilisation de modèles hiérarchiques latents (HLCM, Hierarchical Latent Class Model) (Nefian, 2006) et de réseaux bayésiens à deux couches (une couche de variables observées et une couche de variables latentes) (Zhang *et al.*, 2009). Notamment, les modèles hiérarchiques latents sont prometteurs grâce à leur capacité intrinsèque à réduire la dimension des données de façon graduelle. Cependant peu de travaux ont été réalisés dans ce domaine. Le

passage à l'échelle de ces modèles demeure un important problème.

Dans cet article, nous proposons l'emploi de forêts de modèles hiérarchiques latents (FHLCM) afin de réduire la dimension des données. Ces données pourront ensuite être analysées afin d'identifier les marqueurs génétiques impliqués dans la maladie. Fondamentallement, les variables latentes capturent l'information portée par les marqueurs interdépendants sous-jacents. Ces variables latentes peuvent former des clusters de variables dépendantes. Si ces clusters sont appropriés alors les variables latentes pourront être résumées à leur tour par de nouvelles variables latentes, plus générales. Répéter ce processus engendre la formation d'une structure hiérarchique. Premièrement, l'avantage de cette approche repose sur l'emploi des variables latentes pour l'analyse ultérieure des données ; les données étant ainsi réduites. Les FHLCM peuvent être perçus alors comme un outil de data mining. En effet, grâce à la structure hiérarchique du modèle, l'utilisateur peut commencer par les niveaux les plus élevés du modèle, puis en descendant de niveau, il peut "zoomer" sur des régions d'intérêt. L'utilisateur dispose en fait de plusieurs niveaux de réduction de dimension des données. Il peut ainsi déployer, par exemple, une stratégie descendante lors de la recherche d'association SNP-maladie : les tests d'association avec la variable indicatrice sain/malade peuvent être d'abord réalisés avec les variables latentes des plus hauts niveaux, puis lorsque certaines régions à fortes associations sont ciblées, l'utilisateur peut descendre de niveau progressivement afin d'idenfitier de plus en plus finement le ou les SNP associés à la maladie. Deuxièmement, dans la problématique d'identification des mutations causales, les FHLCM devraient permettre de distinguer les SNP directement associés (vrais positifs), i.e. les marqueurs causaux, des SNP indirectement associés à la maladie par le LD (faux positifs). Pour cela, des tests d'association SNP-maladie conditionnellement à la variable parente (latente) permettraient de distinguer l'influence d'un SNP de celle des autres SNP présents dans le FHCLM.

A notre connaissance, aucun algorithme n'a été conçu spécifiquement pour l'apprentissage des FHLCM, contrairement aux HLCM. La plupart des algorithmes dédiés à l'apprentissage des HLCM ne parviennent pas à surmonter la dimensionnalité des données lorsque ces dernières dépassent des milliers, voire des centaines de milliers de variables. La contribution de cet article est double : (i) un nouveau cadre d'étude des dépendances entre marqueurs génétiques à l'aide de FHLCM est présenté, (ii) un algorithme supportant le passage à l'échelle, CFHLC, a été conçu pour l'apprentissage de la structure et des paramètres des FHLCM. Suivant la méthode hiérarchique de Hwang et de ses collaborateurs développée pour l'étude des données d'expression génique (Hwang *et al.*, 2006), notre approche apporte deux améliorations : (i) une modélisation plus flexible donc plus proche de la réalité, (ii) un contrôle de la perte d'information liée à la réduction de dimension.

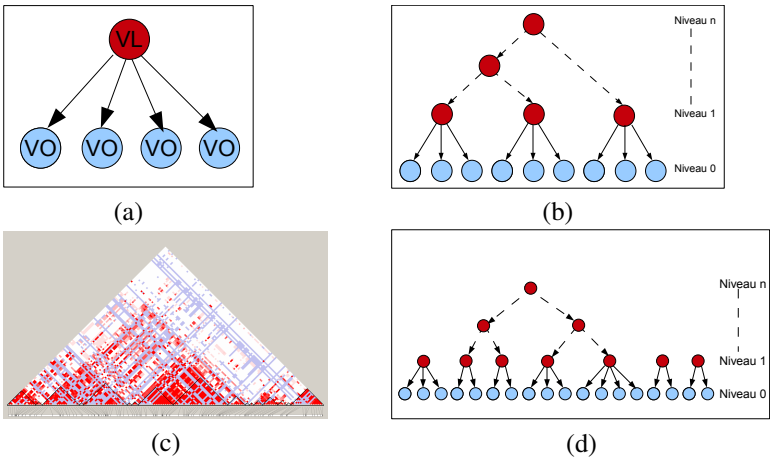
2. Etat de l'art sur l'apprentissage des HLCM

Dans la suite de l'article, nous nous restreignons aux variables finies discrètes (latentes ou observées). Les réseaux bayésiens sont des modèles graphiques probabilistes (Naïm *et al.*, 2007). Ils sont définis par un graphe orienté sans circuit (la structure S) représentant les relations de dépendance au sein de l'ensemble des variables étudiées et par une distribution de probabilités conditionnelles associée à chaque variable (les paramètres Θ). Les modèles latents (LCM, Latent Class Model) forment une classe particulière de réseaux bayésiens : toutes les variables observées (VO) sont dépendantes d'une unique variable latente (VL) (Figure $1(a)$). Les modèles latents sont généralement utilisés pour la classification non supervisée. Cependant, ces modèles reposent sur une hypothèse souvent fausse : l'indépendance locale (Zhang *et al.*, 2004), c'est-à-dire que les VO sont toutes mutuellement indépendantes conditionnellement à la VL. Les HLCM généralisent les modèles latents et ne se basent plus sur cette hypothèse. Leur structure est celle d'un arbre dont les feuilles sont des VO et les noeuds internes sont des VL (par exemple Figure $1(b)$). Les HLCM sont constitués de plusieurs couches (nommées aussi niveaux).

Comme pour les autres réseaux bayésiens, l'apprentissage de la structure et des paramètres des HLCM est nécessaire. Généralement, l'apprentissage de la structure est la tâche la plus difficile à cause de la complexité de l'espace de recherche. Cette tâche peut être classée en deux catégories. La première, l'espérance maximisation structurelle (SEM) optimise alternativement $\Theta|S$ et $S|\Theta$. Dans cette optique, une recherche gloutonne optimisant une fonction de score, telle que le score BIC (Akaike, 1970), a été développée (Zhang, 2003) : l'espace des HLCM est exploré à l'aide d'opérateurs d'addition ou de suppression de VL ou d'états aux variables déjà présentes dans le modèle. D'autres auteurs ont adapté un algorithme SEM combiné avec un algorithme de recuit simulé pour l'apprentissage de réseaux bayésiens à deux couches (une couche de variables observées et une couche de variables latentes) (Zhang *et al.*, 2009) ; l'algorithme de recuit simulé permettant de réduire le risque de tomber dans un optimum local de vraisemblance. L'approche alternative à l'algorithme SEM implémente une classification ascendante hiérarchique (AHC). Se basant sur une mesure de dépendance par paire entre variables, Wang et ses collaborateurs construisent un arbre binaire dirigé ; ensuite, ils appliquent des étapes de régularisation et de simplification de façon à ce que plus de deux noeuds puissent être les enfants d'une VL (Wang *et al.*, 2006). Hwang et ses collègues, quant à eux, réduisent l'espace de recherche des HLCM aux arbres binaires augmentés par de possibles connexions entre frères (noeuds partageant le même parent à l'intérieur d'une même couche intermédiaire) (Hwang *et al.*, 2006). De plus, ils restreignent les VL à être binaires. Malgré ces restrictions, la dernière approche est la seule dont nous savons qu'elle est capable de traiter des données de grandes dimensions. Dans une application mettant en jeu des données de puces à ADN, plus de 6000 gènes ont été analysés sur 60 individus. A notre connaissance, aucun temps d'exécution n'a été

rapporté concernant cette étude.

Néanmoins, la double restriction binaire (arbre binaire et VL binaires) ainsi que le manque de contrôle concernant la perte d'information lorsque la couche augmente sont de sérieux inconvénients pour réaliser une modélisation suffisamment réaliste et une étude d'association ultérieure avec un minimum de puissance. En outre, les dépendances entre SNP devrait être mieux modélisées à l'aide d'une forêt de HLCM de hauteurs variées permettant de prendre en compte l'existence d'indépendance entre marqueurs éloignés (Figure $1(c)$ et Figure $1(d)$). En effet, un des avantages des forêts de HLCM tient au fait qu'elles ne contraignent pas toutes les variables à être dépendantes les une des autres, que ce soit directement ou indirectement.



(a) (b)

(c) (d)

Figure 1. *(a) Exemple de modèle latent. Les variables observées sont colorées en bleu (couleur claire), tandis que les variables latentes sont colorées en rouge (couleur foncée). (b) Exemple de modèle hiérarchique latent. (c) LD plot (matrice des dépendances statistiques entre paires de marqueurs génétiques). Génome humain, chromosome 2, région de 50kb [234 357kb - 234 407kb]. (d) Exemple de forêt de modèles hiérarchiques latents.*

3. Construction du FHLCM

3.1. *Principe*

Notre méthode prend comme entrée une matrice D_X définie sur un ensemble discret fini, noté $\{0,\ 1,\ 2\}$ pour les SNP, décrivant n individus pour p variables $(X\ =\ X_1, ..., X_p)$. L'algorithme CFHLC renvoie un FHLCM. Un FHLCM est composé d'un graphe orienté sans circuit (DAG), appelé aussi la structure, dont les

composantes connexes représentent des arbres, et de θ, les paramètres d'un ensemble de distributions *a priori* et de distribution localement conditionnelles permettant la définition d'une distribution jointe de probabilité. Deux espaces de recherche sont explorés : l'espace des forêts orientées et l'espace de probabilité. En outre, H, l'ensemble des VL ainsi que la matrix de données associée sont des sorties.

Afin de manipuler des données en grande dimension, notre proposition combine deux stratégies. La première consiste à découper les données du génome en régions contigües. Dans notre situation, le découpage en fenêtres de grande taille n'est pas abusif ; il se base sur des propriétés biologiques : la vaste majorité des dépendances entre marqueurs génétiques sont observées entre SNP proches. Ensuite, un FHLCM est appris à l'intérieur de la fenêtre en cours. A l'intérieur d'une fenêtre, une procédure AHC permet de construire itérativement la forêt à partir de deux étapes successives et complémentaires : (i) à chaque étape d'agglomération, une méthode de partitionnement est employée afin d'identifier les clusters de variables ; (ii) chaque cluster est susceptible d'être synthétisé par une VL, à travers un LCM. Pour chaque LCM, l'apprentissage des paramètres et l'imputation des données manquantes (des VL) sont effectués.

3.2. *Partitionnement des noeuds*

Suivant Martin et VanLehn (Martin *et al.*, 1995), idéalement, dans le graphe non orienté des relations de dépendances entre les variables, nous pourrions proposer d'associer une VL à chaque clique présentant des dépendances par paire entre variables. Cependant, la recherche de telles cliques est une tâche NP-difficile. De plus, contrairement à l'objectif des auteurs précédents, les FHLCM n'autorisent pas les noeuds à posséder plus de deux parents : des clusters non-chevauchants de variables sont donc nécessaires pour notre problématique. Ainsi, une méthode approximative résolvant un problème de partitionnement en cliques de variables et prenant en entrée des mesures de dépendances par paire apparait comme une solution efficace et rapide.

3.3. *Apprentissage des paramètres et imputation des données manquantes*

Une étape difficile est de choisir - idéalement d'optimiser - la cardinalité de chaque VL. Au lieu d'employer une valeur arbitraire constante pour chaque VL, nous proposons d'estimer la cardinalité de chaque VL à l'aide d'une fonction du nombre de variables présentes dans le cluster. A chaque étape de la procédure AHC, l'apprentissage des paramètres doit être effectué pour autant de LCM qu'il y a de clusters d'au moins deux noeuds. Pour chaque LCM, l'apprentissage des paramètres peut être réalisé à l'aide d'une procédure EM standard. Cette procédure prend en entrée la cardinalité de la VL et renvoit la distribution de probabilité : les distributions *a priori* pour les VL et les distributions conditionnelles pour les noeuds restants. Une fois les distributions estimées, une façon d'imputer les valeurs manquantes peut consister à

les inférer directement par inférence probabiliste. Finalement, de nouvelles données sont disponibles pour alimenter une nouvelle étape de la construction du FHLCM : les VL découvertes lors de l'étape i seront ensuite considérées comme des variables observées durant l'étape suivante $i + 1$.

3.4. *Contrôle de la perte d'information*

Contrairement à l'approche de Hwang et de ses collaborateurs, dont l'objectif est la compression de données, le contrôle de la perte d'information est nécessaire dans notre cas : chaque VL candidate H à l'étape i ne portant pas suffisamment d'information sur ces noeuds enfants doit être invalidée. En conséquence, ces enfants seront considérés comme des noeuds isolés (*i.e.* des clusters d'une seule variable) à l'étape $i + 1$. Le critère d'information employé, $\mathcal{C}$, repose sur l'information mutuelle moyenne. Il est normalisé grâce à l'entropie $\mathcal{H} : \mathcal{C} = \frac{1}{s_H} \sum_{i \in cluster(H)} \frac{\mathcal{I}(X_i,H)}{\min(\mathcal{H}(X_i), \mathcal{H}(H))}$, avec s_H la taille du $cluster(H)$.

4. Sketch de l'algorithm CFHLC

Pour des raisons d'espace, dans cet article, nous ne présenterons que le sketch de l'algorithme CFHLC. La justification de l'estimation de la cardinalité des VL, aussi bien que les détails à propos de la méthode d'imputation employée pour l'apprentissage des paramètres des LCM est présentée plus en détails dans (Mourad *et al.*, 2010). Pour l'exécution de CFHLC, l'utilisateur dispose de plusieurs paramètres : s, la taille de la fenêtre, spécifie le nombre de SNP contigus (*i.e.* variables) à l'intérieur de celle-ci ; t a pour but de contraindre la perte d'information à un seuil minimal. Les paramètres a, b et $card_{max}$ participent à l'estimation de la cardinalité de chaque VL. Finalement, le paramètre *AlgoPartitionnement* apporte une flexibilité au niveau du choix de la méthode employée pour classifier de manière non-supervisée les variables fortement corrélées à l'intérieur de clusters non-chevauchants.

A l'intérieur de chaque fenêtre successive, la procédure AHC est initiée à partir de la première couche de modèles univariés. Un modèle univarié est construit pour chaque VO de l'ensemble W_i (lignes 4 à 6). La procédure AHC arrête lorsque plus aucun cluster n'est identifié pour synthétiser l'information (ligne 10) ou lorsque plus aucun cluster de taille strictement supérieure à 1 n'a pas été validé (ligne 23). Chaque cluster d'au moins deux noeuds est sujet à l'apprentissage du LCM correspondant, suivi d'une étape de validation (ligne 13 à 22). Afin de simplifier l'apprentissage du FHLCM, la cardinalité des VL est estimée à l'aide d'une fonction affine du nombre de variables du cluster correspondant (ligne 14). L'algorithme *ApprentissageLCM* est pluggé à l'intérieur d'un cadre générique (ligne 15). Après validation à l'aide d'un seuil t (lignes 16 et 17), le LCM est utilisé pour enrichir le FHLCM associé à la fenêtre en cours (ligne 18) : (i) un processus spécifique de fusion lie le noeud addi-

tionnel correspondant à la VL à ses noeuds enfants ; (ii) les distributions *a priori* des noeuds enfants sont remplacées par des distributions conditionnelles à la VL. Dans la fenêtre W_i, les clusters de variables sont remplacés par les VL correspondantes ; la matrice de données $D[W_i]$ est mise à jour en même temps (lignes 19 et 20). *A contrario*, les noeuds des clusters invalidés sont conservés comme noeuds isolés pour l'étape suivante. A la fin, la collection de forêts (DAG) est successivement augmentée par chaque forêt construite à l'intérieur de la fenêtre (ligne 26). En parallèle, sous l'hypothèse d'indépendance entre fenêtres, la distribution jointe du FHLCM final est simplement calculée comme le produit des distributions associées à chaque fenêtre (ligne 26).

Algorithme $CFHLC(X, D_X, s, t, AlgoPartitionnement, a, b, c_{max})$

ENTRÉE **:**
$\mathbf{X}, \mathbf{D_X}$: un ensemble de p variables $X = X_1, ..., X_p$ et la matrice de données correspondante observée sur n individus,
$\mathbf{s}$: une taille de fenêtre,
$\mathbf{t}$: un seuil utilisé pour limiter la perte d'information lors de la construction du FHLCM,
$\mathbf{a}, \mathbf{b}, \mathbf{c_{max}}$: les paramètres utilisés pour calculer la cardinalité des VL.

SORTIE **:**
DAG, θ : le DAG et les paramètres du FHLCM construit,
H, D_H : l'ensemble des VL identifiées lors de la construction du modèle ($H = \{H_1, ..., H_m\}$) et la matrice de données correspondante imputée sur n individus.

$nbFenetres \leftarrow p/s$;
$DAG \leftarrow \emptyset$; $\theta \leftarrow \emptyset$; $H \leftarrow \emptyset$; $D_H \leftarrow \emptyset$
pour $i = 1$ **à** $nbFenetres$
  $W_i \leftarrow \{X_{(i-1)\times s+1}, ..., X_{i\times s}\}$; $D[W_i] \leftarrow D[(i-1) \times s + 1 : i \times s)]$
  $\{\cup_{j\in W_i} DAG_{univ_j}, \cup_{j\in W_i} \theta_{univ_j}\} \leftarrow ApprentModelUniv(W_i)$
  $DAG_i \leftarrow \cup_{j\in W_i} DAG_{univ_j}$; $\theta_i \leftarrow \cup_{j\in W_i} \theta_{univ_j}$

  $etape \leftarrow 1$
  tant que vraie
    $\{C_1, ..., C_{nc}\} \leftarrow Partitionnement(W_i, D[W_i], AlgoPartitionnement)$
    si tous les clusters de taille 1 **aller à fin si**

    $C_{j_1}, ..., C_{j_{nc_2}} \leftarrow ClustersContenantAuMoinsDeuxNoeuds(C_1, ..., C_{nc})$
    $nc_{2_{valide}} \leftarrow 0$
    pour k = 1 **à** nc_2
      $card_H \leftarrow min(EntierArrondi(a \times NombreDeVariables(C_{j_k}) + b, c_{max})$
      $\{DAG_{j_k}, \theta_{j_k}, H_{j_k}, DH_{j_k}\} \leftarrow ApprentissageLCM(C_{j_k}, D[C_{j_k}], card_H)$
      si $(\mathcal{C}(DAG_{j_k}, D[C_{j_k}] \cup DH_{j_k}) \geq t)$ /* validation du cluster en cours - voir Section 3.4 */
        $incr(nc_{2_{valide}})$
        $DAG_i \leftarrow FusionDags(DAG_i, DAG_{j_k})$; $\theta_i \leftarrow FusionParams(\theta_i, \theta_{j_k})$
        $H \leftarrow H \cup H_{j_k}$; $D_H \leftarrow D_H \cup DH_{j_k}$
        $D[W_i] \leftarrow (D[W_i] \setminus D[C_{j_k}]) \cup DH_{j_k}$; $W_i \leftarrow (W_i \setminus C_{j_k}) \cup H_{j_k}$
      fin si
    fin pour
    si $(nc_{2_{valide}} = 0)$ **aller à fin si**
    $incr(etape)$
  fin tant que
  $DAG \leftarrow DAG \cup DAG_i$; $\theta \leftarrow \theta \times \theta_i$
fin pour

Algorithme $ApprentissageLCM(C_r, D[C_r], card_H)$

$H_r \leftarrow CreationNouvelleVariableLatent()$
$DAG_r \leftarrow ConstructionStrucNaive(H_r, C_r)$
$\theta_r \leftarrow StandardEM(DAG_r, D[C_r], card_H)$
$DH_r \leftarrow Imputation(\theta_r, D[C_r])$

Tableau 1. *Sketch de l'algorithm CFHLC.*

5. Résultats expérimentaux et discussion

L'algorithme CFHLC a été implémenté en C++ et repose sur la librairie ProBT dédié aux réseaux bayésiens (http ://bayesian-programming.org). Nous avons pluggé dans CFHLC une méthode de partitionnement conçue par Ben-Dor et ses co-auteurs

(Ben-Dor *et al.*, 1999). CFHLC a été lancé sur un ordinateur standard (3.8 GHz, RAM 3.3 Go). Nous avons généré des données génotypiques simulées en utilisant le logiciel HAPSIMU (http ://l.web .umkc.edu/liujian/). Le paramètre n a été fixé à 2000. Trois tailles d'échantillon sont considérées : $1k$, $10k$ et $100k$ (variables observées). Dans cette article, nous montrons les résultats obtenus avec des paramètres par défaut : $a = 0.2$, $b = 2$, $c_{max} = 20$, $t = 0.5$ (voir Mourad *et al.* (2010) pour de plus amples développements à propos de l'influence des paramètres de CFHLC). La Figure $1(a)$ montre que seulement 15 heures sont nécessaires pour 10^5 SNP, en considérant une taille de fenêtre s de 100. Pour le même jeu de données traité dans les cas "$s = 200$" et "$s = 600$", les temps d'exécution sont de 20.5 h et de 62.5 h, respectivement. Pour le même nombre de variables observées ($100k$), Wang *et al.* rapporte des temps d'exécution de l'ordre de deux mois. La figure $1(b)$ décrit plus précisément l'influence de l'augmentation de la taille de la fenêtre sur le temps d'exécution. De façon intéressante, la Figure $1(c)$ souligne la diminution du nombre de variables à analyser par tests d'association avec la variable indicatrice sain/malade (de 1000 variables observées à moins de 200 racines de la forêt dans le cas "$s = 100$"). Dans le cas précédent, CFHLC permet une réduction du nombre de variables à analyser de plus de 80%. Pour ce même cas, la Figure $1(d)$ révèle une diminution importante du nombre de LV par couche (en considérant le modèle dans son ensemble) lorsque la couche augmente. Puisque la plupart des LV sont présentes dans la première couche (64%), la perte d'information des variables observées est limitée. Finalement, la Figure $1(e)$ montre comment l'information diminue lorsque le nombre de couches augmente. Dans les couches les plus élevées, l'information mutuelle normalisée moyenne est au moins égale à 0.52 et 0.56 pour les cas "$s = 100$" et "$s = 600$", respectivement. Ainsi, bien que CFHLC permette le passage à l'échelle, il assure en même temps un contrôle efficace au niveau de la perte d'information. Plus de résultats et de commentaires associés sont présentés dans Mourad *et al.* (2010).

A notre connaissance, notre méthode hiérarchique est la première à s'être révélée capable de traiter rapidement l'apprentissage de modèle pour des données pangénomiques. Alors que l'objectif de Hwang et de ses collaborateurs est la compression de données, nous sommes face à un problème plus difficile : permettre une étude d'association en aval tout en conservant une puissance suffisante. Le relâchement de la double restriction binaire (arbre binaire et VL binaire) fournit un cadre d'étude intéressant pour les études d'association pangénomiques : en particulier, la flexibilité de la taille des clusters permet de réduire le nombre de VL à l'intérieur du modèle.

6. Conclusion

Notre contribution dans cet article est double : (i) un nouveau cadre d'étude des dépendances entre marqueurs génétiques à l'aide de FHLCM est présenté ; (ii) CFHLC, un algorithme dédié à apprendre ces modèles, s'est révélé capable de traiter efficacement des données pangénomiques lors de benchmarks. Concernant le partitionnement des noeuds et l'imputation des VL, un de nos travaux en cours consiste à examiner

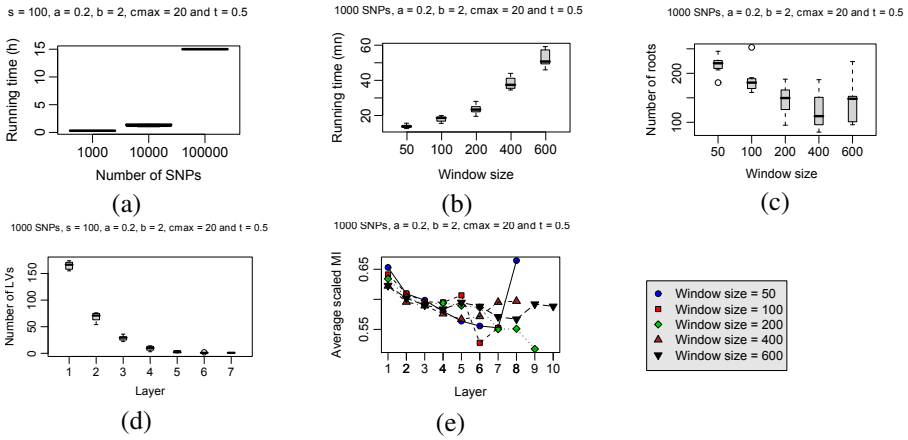


Figure 1 : (a) Temps d'exécution *versus* nombre de variables. (b) Temps d'exécution *versus* taille de la fenêtre. (c) Nombre de racines *versus* taille de la fenêtre. (d) Nombre de VL par couche pour l'ensemble du FHLCM. (e) Effet de la taille de la fenêtre sur l'information mutuelle normalisée moyenne par couche pour l'ensemble du FHLCM. N.B. : les boîtes à moustache sont réalisées pour 20 benchmarks (exceptionnellement 5 dans le cas $100k$ (a)).

quelle méthode à plugger est la plus pertinente, plus particulièrement pour les données de GWAS. Finalement, nous évaluerons l'algorithme CFHLC pour d'autres applications que la réduction de dimension des données pangénomiques comme l'étude et la visualisation du déséquilibre de liaison, la cartographie des SNP causaux et l'étude de la structure de la population.

7. Bibliographie

Akaike H., « Statistical Predictor Identification », *Ann. Inst. Statist. Math.*, vol. 22, p. 203-217, 1970.

Ben-Dor A., Shamir R., Yakhini Z., « Clustering Gene Expression Patterns », *Proceedings of the third annual international conference on Computational molecular biology*, p. 33-42, 1999.

Hwang K.-B., Kim B.-H., Zhang B.-T., « Learning hierarchical bayesian networks for large-scale data analysis », *ICONIP*, p. 670-679, 2006.

Martin J., Vanlehn K., Discrete factor analysis : Learning hidden variables in bayesian network, Technical report, Department of Computer Science, University of Pittsburgh, 1995.

Mourad R., Sinoquet C., Leray P., « Learning a forest of Hierarchical Bayesian Networks to model dependencies between genetic markers », , LINA, Research Report, hal-00444087, 2010.

Naïm P., Wuillemin P.-H., Leray P., Pourret O., Becker A., *Réseaux bayésiens*, 3 edn, 2007.

Nefian A. V., « Learning SNP dependencies using embedded Bayesian networks », *IEEE Computational Systems, Bioinformatics Conference*, 2006.

Pattaro C., Ruczinski I., Fallin D. M., Parmigiani G., « Haplotype block partitioning as a tool for dimensionality reduction in SNP association studies », *BMC Genomics*, vol. 9, p. 405, August, 2008.

Schaid D. J., « Evaluating association of haplotypes with traits », *Genetic Epidemiology*, vol. 27, p. 348-364, 2004.

Stram D. O., « Tag SNP selection for association studies », *Genetic Epidemiology*, vol. 27, p. 365-374, 2004.

Verzilli C. J., Stallard N., Whittaker J. C., « Bayesian graphical models for genomewide association studies », *The american journal of human genetics*, vol. 79, p. 100-112, 2006.

Wang Y., Zhang N. L., Chen T., « Latent Tree Models and Approximate Inference in Bayesian Networks », *Machine Learning*, vol. 32, p. 879-900, 2006.

Zhang N. L., Structural EM for Hierarchical Latent Class Models, Technical report, 2003.

Zhang N. L., Kocka T., « Efficient learning of hierarchical latent class models », *Proceedings of the 16th IEEE International Conference on Tools with Artificial Intelligence (ICTAI)*, p. 585-593, 2004.

Zhang Y., Ji L., « Clustering of SNPs by a Structural EM Algorithm », *International Joint Conference on Bioinformatics, Systems Biology and Intelligent Computing*, p. 147-150, 2009.

Article reçu le 22/09/1996.
Version révisée le 04/10/2005.
Rédacteur responsable : GUILLAUME LAURENT

SERVICE ÉDITORIAL – HERMES-LAVOISIER
14 rue de Provigny, F-94236 Cachan cedex
Tél. : 01-47-40-67-67
E-mail : revues@lavoisier.fr
Serveur web : http://www.revuesonline.com

ANNEXE POUR LE SERVICE FABRICATION
A FOURNIR PAR LES AUTEURS AVEC UN EXEMPLAIRE PAPIER
DE LEUR ARTICLE ET LE COPYRIGHT SIGNE PAR COURRIER
LE FICHIER PDF CORRESPONDANT SERA ENVOYE PAR E-MAIL

1. ARTICLE POUR LA REVUE :

2. AUTEURS :

Raphaël Mourad[*] *— Christine Sinoquet* [**] *— Philippe Leray* [*]

3. TITRE DE L'ARTICLE :

Apprentissage de réseaux bayésiens hiérarchiques latents pour les études d'association pangénomiques

4. TITRE ABRÉGÉ POUR LE HAUT DE PAGE MOINS DE 40 SIGNES :

r

5. DATE DE CETTE VERSION :

15 mars 2010

6. COORDONNÉES DES AUTEURS :

– adresse postale :
[*] LINA, UMR CNRS 6241, Ecole Polytechnique de l'Université de Nantes,
rue Christian Pauc, BP 50609, 44306 Nantes Cedex 3, France
{raphael.mourad,philippe.leray}@univ-nantes.fr
[**] LINA, UMR CNRS 6241, Université de Nantes,
2 rue de la Houssinière, BP 92208, 44322 Nantes Cedex 3, France
christine.sinoquet@univ-nantes.fr
– téléphone : 00 00 00 00 00
– télécopie : 00 00 00 00 00
– e-mail : guillaume.laurent@ens2m.fr

7. LOGICIEL UTILISÉ POUR LA PRÉPARATION DE CET ARTICLE :

LaTeX, avec le fichier de style `article-hermes.cls`,
version 1.23 du 17/11/2005.

8. FORMULAIRE DE COPYRIGHT :

Retourner le formulaire de copyright signé par les auteurs, téléchargé sur :
`http://www.revuesonline.com`

SERVICE ÉDITORIAL – HERMES-LAVOISIER
14 rue de Provigny, F-94236 Cachan cedex
Tél. : 01 47 40 67 67